Vers une démocratisation des outils de constitution de corpus parallèles

Octavia Efraim[#1], Fabienne Moreau[#2]
LIDILE EA 3874, Université Rennes 2, Place du recteur Henri Le Moal, CS 24307, 35043 RENNES cedex
#1-octavia-edie.efraim@uhb.fr, #2-fabienne.moreau@uhb.fr

Résumé. Si la traduction automatique (TA) a désormais conquis sa place dans le milieu de la traduction, que ce soit dans un contexte de formation (*e.g.* l'apparition de cours de post-édition) ou en milieu professionnel (*e.g.* l'intégration des outils de TA directement à l'environnement de TAO), l'étape cruciale consistant à personnaliser ces outils est encore aujourd'hui hors de portée du traducteur. En effet, les corpus bilingues disponibles sont rares et souvent peu adaptés car non spécialisés, et les outils existants pour constituer de telles ressources sont trop complexes à utiliser pour les (apprentis-)traducteurs. Ce travail vise à démocratiser la constitution de ces ressources parallèles. Dans le cadre d'une activité pédagogique, nous avons proposé de simplifier au maximum la procédure nécessaire à l'élaboration des corpus et de constituer une boîte à outils permettant d'enchaîner plus facilement les tâches du processus. Une automatisation plus poussée du processus est également envisagée.

Abstract. Machine translation (MT) has won its place in the world of translation: MT-related contents (such as post-editing) are now a fixture in the translation curriculum, and in professional settings MT is accessed through plug-ins within CAT environments. Nonetheless, MT engine customisation – a crucially important task for an MT system's performance – remains too often out of translators' reach. Indeed, bilingual corpora are rarely available, and often ill-suited to the task (few are domain-specific). Moreover, the tools available to (trainee) translators for building training corpora are still too complex for them to use. Our work aims at democratising such tools. As part of a hands-on activity, we set out to simplify the parallel corpus building process, by assembling a 'toolbox' which handles the process as a sequence of easier-to-handle tasks. Further automation of the process is possible.

Mots-clés : corpus parallèle, corpus d'entraînement, corpus bilingue, traduction automatique statistique (TAS)

Keywords: parallel corpus, training corpus, bilingual corpus, statistical machine translation (SMT)

1 Motivation et contexte

Désormais incontournable sur le marché de la traduction, la TA statistique (TAS) reste sous-exploitée par le traducteur, qui l'utilise souvent dans sa version « grand public » ou déjà entraînée. Dans les cursus de traduction, elle fait cependant l'objet d'activités pédagogiques qui mettent les apprentis-traducteurs en situation de personnaliser des moteurs de TAS. L'acquisition des bitextes nécessaires à ce processus n'est toutefois généralement pas le focus de ces projets étudiants, qui se limitent souvent à exploiter des corpus déjà existants (Doherty, Kenny, 2014), telles les ressources bilingues de l'UE, qui, malgré leur qualité, restent limitées quant aux domaines couverts. Or nous croyons que la capacité du traducteur à gérer et optimiser en toute autonomie les outils de TAS passe nécessairement par la maîtrise de l'acquisition de corpus bilingues spécialisés.

En effet, si la gestion des outils de TAS est désormais accessible aux non-techniciens, la quasi-absence de corpus parallèles spécialisés en libre accès et le prix élevé des ressources payantes freinent l'adoption à plus large échelle de cette technologie, les (apprentis-)traducteurs qui souhaitent recourir à des systèmes de TAS étant contraints de constituer eux-mêmes ces textes parallèles. Or, dans l'état de l'art actuel, cette tâche s'avère ardue et est souvent réservée aux concepteurs (informaticiens) des systèmes de TA, car les outils et les plateformes existants sont peu adaptés puisque :

1. les outils disponibles au grand public (peu nombreux, la plupart étant des prototypes de recherche ; citons Bitextor[1] (Esplà-Gomis, 2009) ou ILSP-FC[2] (Papavassiliou *et al.* 2013)) requièrent une procédure

[1] http://sourceforge.net/projects/bitextor/

[2] http://nlp.ilsp.gr/redmine/projects/. Des flots de travail intégrant ce dernier sont disponibles à http://myexperiment.elda.org/workflows/37 ou encore http://myexperiment.elda.org/workflows/7.

d'installation complexe, voire un système d'exploitation spécifique. S'ensuivent, pour le non-technicien, une perte de temps importante et un renoncement lié à la complexité du processus ;

2. les outils « clé en main », gérant l'intégralité du processus (recherche et récupération de paires de documents pertinents issus du Web, nettoyage, renommage, formatage et alignement des documents) sont rares et réservés aux chercheurs (*e.g.* STRAND et BITS (Tiedemann, 2011), ou, plus récemment, PaCO[2] (San Vicente, Manterola, 2012) ou ILSP-FC). Il faut souvent multiplier les outils et compléter les opérations (semi-)automatiques par des traitements manuels, qui peuvent être lourds et complexes si le volume de textes à intégrer est conséquent. Or, la valeur ajoutée du traducteur par rapport à l'informaticien résidant dans sa capacité à évaluer la qualité des textes, c'est à la gestion de l'aspect linguistique de la tâche, et non de celui technique, qu'il devrait essentiellement consacrer son temps.

Dans ce contexte, nous proposons une « boîte à outils » que nous avons compilée pour permettre à nos étudiants en traduction de réaliser une activité pédagogique liée à la TA, consistant à entraîner des moteurs de TAS (anglais > français) pour différents domaines[3]. Sans se prétendre innovante, notre approche présente la particularité d'être simple et accessible à un public non-informaticien. La solution proposée décompose le processus d'acquisition du corpus bilingue en étapes clairement délimitées, et vise à limiter autant que possible le nombre de tâches effectuées. Après avoir présenté le mode d'emploi simple élaboré pour constituer aisément des corpus parallèles, nous évoquerons les principales difficultés qui peuvent être rencontrées et suggérerons des perspectives d'automatisation plus poussées.

2 Solution proposée

Les traitements proposés se décomposent selon les opérations décrites dans le tableau 1. Nous avons cherché à intégrer soit des logiciels en libre accès, soit des outils pour lesquels un traducteur peut détenir déjà une licence d'utilisation.

Étape	Opération	Type d'outil	Outil(s) utilisé(s) dans notre expérimentation	Améliorations envisagées
1.	Constitution d'une liste de sites Web bilingues spécialisés	Moteur de recherche Web + opérateurs de recherche	Google + inurl:/english/, site:.ca, etc.	Couplage du formulaire de recherche avec l'aspiration des sites
2.	Téléchargement des sites Web retenus	Aspirateur de sites Web	HTTrack[4]	
3.	Regroupement des fichiers parallèles dans des dossiers parallèles/dans le même dossier (selon les exigences de l'aligneur)	Opération manuelle/partiellement automatisable via un script « maison »	-	Couplage des deux tâches et automatisation aussi poussée que possible par script
4.	Renommage des fichiers parallèles selon le format exigé par l'aligneur	Opération manuelle/partiellement automatisable via un logiciel de renommage, ou un script « maison »	-/Bulk Rename Utility[5]	

[3] Nous avons opté délibérément pour des domaines pour lesquels des corpus parallèles gratuits ne sont pas disponibles : les assurances, l'habillement, le tourisme et l'hôtellerie.

[4] http://www.httrack.com/

[5] http://www.bulkrenameutility.co.uk/

(A) Si pré-alignement non requis par le système de TA[6] :				
A5.	Préparation des fichiers pour leur chargement dans le système : selon les exigences du système de TAS, découpage des fichiers et des dossiers en blocs ne dépassant pas la taille maximale admise	Opération manuelle/partiellement automatisable via un logiciel de découpage de dossiers	Folder Axe[7]	Gestion du découpage automatique en mode récursif sans destruction de la hiérarchie de dossiers
(B) Si pré-alignement requis :				
B5.	Alignement des fichiers parallèles	Aligneur	AlignFactory/aligneur de memoQ	
B6.	Pré-traitement des bitextes : élimination d'éléments parasites, uniformisation terminologique, etc.	Script « maison »/tableur	Script/Excel	
B7.	Correction et validation des bitextes	Opération manuelle	Éditeur d'alignement (AlignFactory/memoQ)	
B8.	Identique à A5.			

Tableau 1 : Procédure proposée pour l'acquisition de corpus parallèles

La chaîne de traitements comporte un minimum de quatre et un maximum de huit étapes. Les tâches 1-4 sont indispensables quelles que soient les exigences du système de TAS auquel le corpus est destiné. L'étape B6 n'est applicable que si le format de sortie de B5 le permet sans compromettre B7, d'éventuelles conversions de formats étant alors également à prévoir.

3 Résultats

Les étudiants ont trouvé les étapes du processus claires et ont réussi à constituer des corpus parallèles et à entraîner des moteurs de TA. Des améliorations sont envisageables (voir quatrième colonne du tableau 1) pour résoudre certaines des difficultés qu'ils ont signalées, notamment : variabilité de la productivité de la recherche selon le domaine et le couple de langues, surtout lorsqu'on applique des critères de recherche restrictifs (variété géographique de langue, etc.) ; temps de téléchargement important et parfois échec de l'aspiration du site ; traitements manuels aux étapes 3 et 4 largement dépendants de la structure de chaque site ; validation manuelle des alignements très prenante.

4 Conclusion

Dans un contexte de démocratisation des solutions de TAS, l'acquisition des ressources parallèles nécessaires à l'entraînement de ces systèmes reste une tâche encore réservée aux concepteurs. Or ce n'est qu'en mettant à portée de

[6] Dans notre projet, nous utilisons le système de TAS Microsoft Translator (et son interface https://hub.microsofttranslator.com) qui fournit un aligneur pré-intégré. Ce système impose des restrictions sur la taille des documents à télécharger, d'où les étapes A5 et B8.

[7] http://bkprograms.weebly.com/folder-axe.html

tous la possibilité de créer des bitextes de taille importante à partir du Web (Smith et al., 2013) qu'on réussira à rendre réellement autonomes les utilisateurs des technologies de TAS.

Références

DOHERTY S., KENNY D. (2014). The design and evaluation of a Statistical Machine Translation syllabus for translation students. The Interpreter and Translator Trainer 8, 295–315.

ESPLÀ-GOMIS M. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites. Proceedings of MT Summit XII.

PAPAVASSILIOU V., PROKOPIDIS P., THURMAIR G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. Proceedings of the 6th Workshop on Building and Using Comparable Corpora, 43–51.

SAN VICENTE I., MANTEROLA I. (2012). PaCo^2: a fully automated tool for gathering parallel corpora from the Web. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).

SMITH J. R., SAINT-AMAND H., PLAMADA M., KOEHN P., CALLISON-BURCH C., LOPEZ A. (2013). Dirt cheap Web-scale parallel text from the Common Crawl. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 1374-1383.

TIEDEMANN J. (2011). Bitext alignment. Morgan & Claypool.